A STOCHASTIC REPRESENTATION OF THE DYNAMICS OF SUNG MELODY

Yasunori OHISHI
Graduate School of Information Science, Nagoya University
ohishi@sp.m.is.nagoya-u.ac.jp

Masataka GOTO
National Institute of Advanced Industrial Science and Technology (AIST)
m.goto@aist.go.jp

Katunobu ITOU
Faculty of Computer and Information Sciences, Hosei University
itou@k.hosei.ac.jp

Kazuya TAKEDA
Graduate School of Information Science, Nagoya University
kazuya.takeda@nagoya-u.jp

ABSTRACT

In this paper, we propose a stochastic representation of a sung melodic contour, called *stochastic phase representation (SPR)*, which can characterize both musical-note information and the dynamics of singing behaviors included in the melodic contour. The SPR is constructed by fitting probability distribution functions to F0 trajectories in the F0-ΔF0 phase plane. Since fluctuations in singing can be easily separated by using SPR, we applied SPR to a melodic similarity measure for query-by-humming (QBH) applications. Our experimental results showed that the SPR-based similarity measure was superior to a conventional dynamic-programming-based method.

1 INTRODUCTION

The goal of this study is to build a model that can represent the dynamics of various singing behaviors (e.g., fluctuations in a musical note and continuous transitions between notes) in a sung melodic contour. Although a symbolic melodic contour (a sequence of musical notes) can be easily modeled by a discrete-time stochastic representation such as n-grams, this representation cannot be used for modeling a sung melody because it is difficult to represent the singing dynamics of its melodic contour, such as vibrato and overshoot. The dynamic representation for modeling a sung melody is important for defining an appropriate melodic similarity between sung melodies, which is useful for various applications such as query-by-humming (QBH) and automatic clustering of songs.

Most previous studies including symbolic melodic similarities [1, 2] and melodic similarities for sung melodies [3, 4, 5, 6] focused on the retrieval performance. For sung melodies, for example, a melodic contour was represented by a discrete symbolic sequence of musical notes [3, 4] or a sequence of pitch histograms for unstable pitch contours [5, 6]. Since they did not model the dynamics at all, their melodic similarities are sometimes too sensitive to singing behaviors that may differ among singers.

Therefore, we propose a novel stochastic graphical representation of the dynamic properties of sung melodic contours, called *stochastic phase representation (SPR)*. This representation is a generative model of melodic contours and can separate the dynamics of various singing behaviors from an original musical note sequence. By using this representation, we also define a melodic similarity measure for QBH applications. In our experiments, we show the effectiveness of this similarity measure based on SPR.

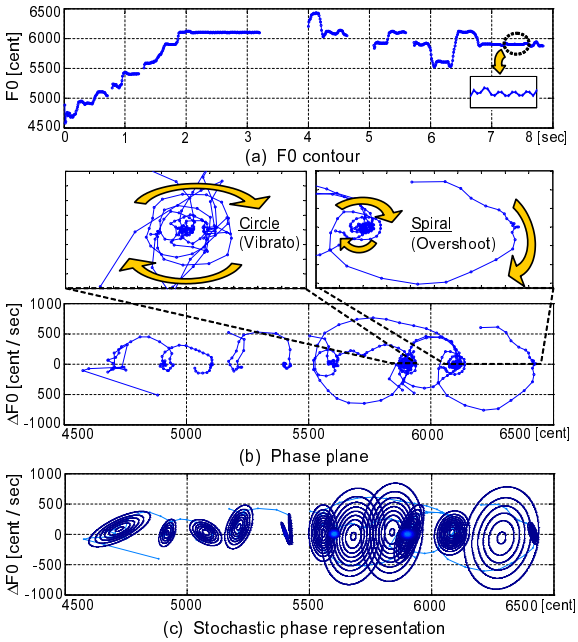

Figure 1. Schematic view of constructing stochastic phase representation (SPR). The original F0 contour (a) is mapped onto the F0-ΔF0 phase plane (b). By fitting Gaussian mixture models to trajectories on the phase plane, stochastic representation of the F0 dynamics (c) can be constructed.

2 STOCHASTIC PHASE REPRESENTATION (SPR) FOR MELODIC CONTOUR

Figure 1 shows an example of an SPR constructed from singing melodic contours represented as trajectories of the fundamental frequency (F0). We assume that the F0 trajectories are generated by a dynamic system and represented in a two-dimensional phase plane, $\vec{f}(x, \dot{x})$, where x is the F0 and $\dot{x}$ is its differential. That is, $\vec{f}(x, \dot{x})$ represents the local direction of an F0 trajectory. A fluctuation in a sung melody can be modeled by a damped oscillation of the dynamic system and appears as a curling trajectory around a certain target point, i.e., an attractor of the system. The advantage of this modeling is that typical singing behaviors can be characterized by the shape of curling trajectories. As shown in Fig. 1(b), for example, a vibrato within a musical note appears as a circular pattern because it has the quasi-periodic modulation of the F0, and an overshoot after a note change appears as a spi-

ral pattern because the F0 of the overshoot transitionally exceeds the F0 of a (target) musical note just after the note change. Here, the location of each attractor corresponds to the F0 of its target musical note.

Therefore, we model the curling trajectories by fitting a Gaussian mixture model (GMM) so that the likelihood of observing the given trajectories becomes the maximum. We refer to this GMM-based representation of the F0 trajectories (sung melodic contours) as *stochastic phase representation (SPR)* shown in Fig. 1(c). The F0 of musical notes is represented by the location of the local maxima of the SPR, and the singing behavior of those notes is represented by the shape around the local maxima. Because each (target) note and its relative length in a melodic contour are captured as the location and its height of the corresponding local maximum, respectively, the divergence between GMM-based distributions in the phase plane is expected to be a robust melodic similarity measure that can reduce variations by singing behaviors and focuses on the original (target) melodic information.

3 EXPERIMENTS

The potential of SPR was preliminarily evaluated on a small QBH application. The song database consists of 50 short excerpts from 25 pop songs of the RWC Music Database (RWC-MDB-P-2001) [7]. The average length of those excerpts is 12 s. For query melodies, 75 subjects listened to each of the above 50 excerpts and then sang its melody with lyrics [8]. The number of recorded samples was 3,750 (75 $\times$ 50), but we used 3,257 samples after excluding samples whose melody was extremely different from the original melody. [1]

The F0 contour of the query melodies was estimated for every 10 ms by using YIN [9]. The F0 contour of the 50 excerpts in the song database was manually annotated [10]. Both F0 contours were represented in cents so that one equal-tempered semitone corresponds to 100 cents, and then normalized by subtracting the average F0 value over each contour.

Finally, the similarity between a query melody and each excerpt in the song database was calculated by using a histogram-intersection distance [11] between their discretized SPRs. SPRs were modeled by 16-mixture GMMs and converted into discretized SPRs where F0 and ΔF0 were uniformly partitioned into square cells (100 cent F0 $\times$ 25 cent/sec ΔF0) and relative occurrences (frequencies) within square cells were calculated.

However, since this discretized-SPR-based distance did not take into account the temporal order of notes, we divided a long contour into several short segments so that short segments of the query can be compared with the corresponding short segments of each database excerpt in order. Their similarity was calculated by the cumulative sum of their distances. We thus investigated the performance improvement by increasing the number of segments. As for the baseline performance, we also evaluated a traditional dynamic time warping (DTW) matching technique using F0 contours.

[1] Since all songs in RWC-MDB-P-2001 were original compositions, the subjects were not familiar with these melodies.

Table 1. Percentage of Mean Reciprocal Rank (MRR)

	DTW	Proposed			
# of segment		1	2	4	8
MRR [%]	64.3	45.6	57.1	65.6	71.1

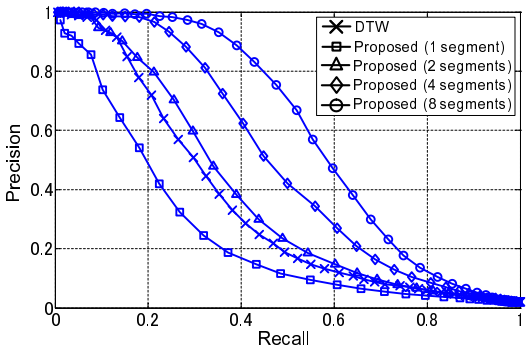

Figure 2. ROC curves of similarity measures.

4 RESULTS AND DISCUSSIONS

The obtained QBH results of the mean reciprocal rank (MRR) and ROC curves are shown in Table 1 and Fig. 2. The proposed distance using the original contours was inferior to the baseline DTW. However, if we used the proposed cumulative distance after dividing each query contour and each database excerpt into eight segments, the MRR performance and the ROC curve were improved and were better than the DTW.

These preliminary results showed that our histogram-based distance using SPR is promising for measuring melodic similarity. In the future, we plan to evaluate it in detail on a larger database. Although SPR has great potential for representing and generating singing dynamics, we have not tested it yet. Future work will include the evaluation of its ability to automatically detect particular singing behaviors such as vibrato and overshoot, and the generation of melodic contours that reflect personal singing behaviors.

5 REFERENCES

[1] Typke, R. et al., "Using transportation distances for measuring melodic similarity," *Proc. ISMIR*, 2003.

[2] Grachten, M. et al., "Melodic Similarity: Looking for a Good Abstraction Level," *Proc. ISMIR*, 2004.

[3] Hu, N. et al., "A probabilistic model of melodic similarity," *Proc. ICMC*, 2002.

[4] Pauws, S., "CubyHum: A fully operational query by humming system," *Proc. ISMIR*, 2002.

[5] Adams, N. H. et al., "Time Series Alignment for Music Information Retrieval," *Proc. ISMIR*, 2004.

[6] Song, J. et al., "Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System," *Proc. ISMIR*, 2002.

[7] Goto, M. et al., "RWC music database: Popular, classical, and jazz music databases," *Proc. ISMIR*, 2002.

[8] Goto, M. et al., "AIST Humming Database:Music Database for Singing Research," *IPSJ (MUS)*, Vol. 2005, No. 82, pp. 7-12, 2005 (in Japanese).

[9] de Cheveigne, A. et al., YIN, "a fundamental frequency estimator for speech and music," *JASA*, Vol.111, No.4, pp.1917-1930, 2002.

[10] Goto, M., "AIST Annotation for the RWC Music Database," *Proc. ISMIR*, 2006.

[11] Kashino, K. et al., "A Quick Search Method for Audio and Video Signals Based on Histogram Pruning," *IEEE Trans. Multimedia*, Vol.5, No.3, pp.348-357, 2003.